Text Alignment

Satz-, Offset- und Wort-Alignment

Caren Brinckmann

12.12.2001

Spezielle statistische Methoden in der Sprachdatenverarbeitung

- Text Alignment: Kapitel 13.1 und 13.2 (Manning & Schütze, 1999)
- Was hat Text Alignment mit maschineller Übersetzung zu tun?
  - Prinzipiell können alle Module eines maschinellen Übersetzungssystems mit statistischen Methoden implementiert werden, obwohl in der Praxis statistische und regelbasierten Ansätze kombiniert werden (vgl. Verbmobil: paralleles Parsing durch einen Chunker, einen probabilistischen Parser und einen HPSG-Parser).
  - Textalignment-Techniken werden verwendet, um Material bereitzustellen, auf das die statistischen MÜ-Module zugreifen (z.B. probabilistische bilinguale Wörterbücher und parallele Grammatiken).
- Text Alignment kann auf verschiedenen Ebenen realisiert werden (die später genauer erläutert werden):
  - Absatz-Alignment
  - Satz-Alignment
  - Offset-Alignment
  - Wort-Alignment

Text Alignment - Übersicht

- Einführung, Anwendungen und Probleme
- Satz-Alignment:
  - Länge-basierte Methode: Gale & Church (1993)
  - Lexikalische Methode: Kay & Röscheisen (1993)
- Offset-Alignment: Church (1993)
- Wort-Alignment

Struktur der Präsentation

- Einführung: Was ist überhaupt Text Alignment, wofür braucht man das, und wo liegen die Probleme?
- Vorstellung und Vergleich verschiedener Methoden zum Satz, Offset- und Wort-Alignment

Text Alignment - Einführung

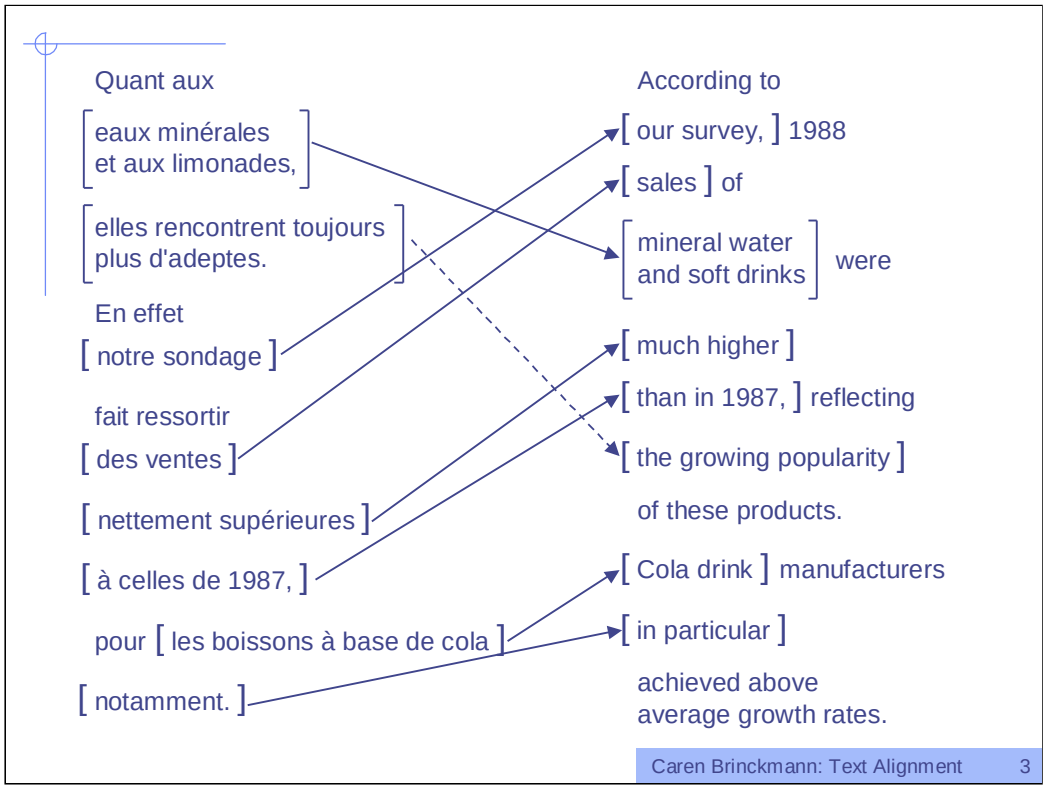
- Text Alignment
  - Anwendung statistischer Methoden auf multilinguale parallele Texte (Korpora)
  - Zuordnung von Absätzen, Sätzen und Wörtern
- Korpora
  - Parlamentsberichte (*Hansards*) und andere offizielle Dokumente aus Ländern mit mehreren Landessprachen (Kanada, Schweiz, Hongkong)
  - Zeitungen, religiöse oder literarische Schriften, die in verschiedenen Sprachen erscheinen
- Verwendungszwecke
  - Probabilistische bilinguale Lexika (Wort Alignment)
  - Maschinelle Übersetzung
  - Wissensbasis für
    - Disambiguierung der Wortbedeutung
    - multilinguales Information Retrieval
  - Hilfsmittel für Übersetzer

- Korpora
  - Vorteile von Parlamentsberichten und anderen offiziellen Dokumenten aus Ländern mit mehreren Landessprachen:
    - in großen Mengen frei verfügbar
    - konsistentes Genre
    - akkurate, wortgetreue Übersetzungen
  - Probleme bei Zeitungen, religiösen oder literarischen Schriften, die in verschiedenen Sprachen erscheinen:
    - unterschiedliche Genres
    - weniger wortgetreue Übersetzung
- Verwendungszwecke
  - Probabilistische bilinguale Lexika

    Beispiel: Auszug aus einem probabilistischem Lexikon (Einträge für das englische Wort „the“, mit der jeweiligen Wahrscheinlichkeit, dass „the“ in das entsprechende französische Wort übersetzt wird)

Englisch	**Französisch**	**P(Franz.|Engl.)**
the	le	0.610
the	la	0.178
the	l’	0.083
the	les	0.023
the	ce	0.013
the	il	0.012
the	de	0.009
the	à	0.007
the	que	0.007

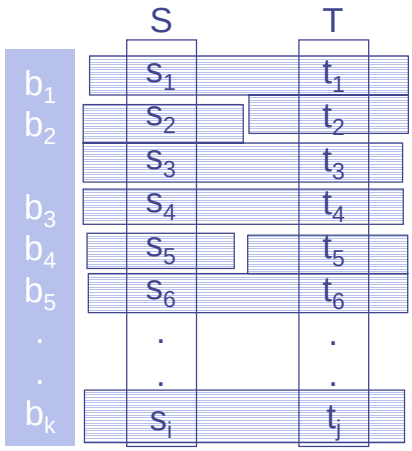
  - Hilfsmittel für Übersetzer
    - Beispiel Bedienungsanleitungen: Wenn eine Bedienungsanleitung aktualisiert wird, müssen nur die geänderten Abschnitte neu übersetzt werden. Um diese Abschnitte herauszufinden, wird Text Alignment verwendet.
    - Übersetzungsdatenbank (Beispiel „European Terminology Database“ http://europa.eu.int/eurodicautom/)



- Ein Beispiel, das verdeutlichen soll, wie sehr Übersetzer das Original verändern bzw. anders anordnen, sogar in eher technischen Texten!
  - Herkunft des Beispiels: Berichte der Union Bank of Switzerland
  - Links: französischer Text, Rechts: englischer Text
  - Die Pfeile verbinden die Satzteile miteinander, die als Übersetzungen voneinander angesehen werden können.
- Auffälligkeiten
  - Der gleiche Sachverhalt wird in beiden Sprachen mit 2 Sätzen dargestellt.
  - Aber: Umstellung von Satzteilen innerhalb eines Satzes und über Satzgrenzen hinweg.
  - „achieved above average growth rates“ hat gar kein Pendant (nur implizit im französischen Text)

Satz-Alignment

- parallele Texte S und T in zwei verschiedenen Sprachen:
  S = (s_1, ..., s_i), T = (t_1,...,t_j)
- jeder Satz aus S und T wird genau einem „*bead*“ (b_1, ..., b_k), d.h. einer Gruppe zugeordnet
- Kriterium: inhaltliche Entsprechung der Sätze innerhalb eines beads
- 1:1 Korrespondenz der Sätze nur zu 90% (auch möglich: 1:2, 2:1, 1:3, 3:1 etc.)
- 1:0 und 0:1 → Einfügungen, Löschungen
- Korrespondenz-Problem: Umstellung der Sätze innerhalb eines Absatzes
  („*crossing dependencies*“ ≈ Kreuzabhängigkeiten)
- Beschränkung auf Alignment **ohne** Kreuzabhängigkeiten → 2:2, 2:3, 3:2 etc.



- Überlappungskriterium: Inwieweit müssen sich Sätze inhaltlich überlappen, damit sie zu einem „bead“ gehören? Ein oder zwei Wörter reichen meist nicht aus, es muss schon mindestens ein Nebensatz sein.
- Beispiel auf Folie 3: 2:2 Korrespondenz

Text Alignment - Methoden

- Länge-basiertes Satzalignment
  - Grundidee: Kurze Sätze werden in kurze Sätze übersetzt, lange Sätze werden in lange Sätze übersetzt.
  - Satzlänge: Anzahl der Wörter vs. Anzahl der Zeichen
  - effizient (durch *Dynamische Programmierung*)
  - erfolgreich bei ähnlichen Sprachen und wortgenauer Übersetzung
- Lexikalische Methoden
  - robustes Satzaligment durch Verwendung lexikalischer Informationen (statistische Verteilung einzelner Wörter)
- Offset Alignment durch Signalverarbeitungstechniken
  - robustes Alignment von Offset Positionen

- Die Satzlänge kann in Anzahl der Wörter oder in Anzahl der Zeichen gemessen werden (vgl. Folie 9)
- *Dynamische Programmierung*: rekursiver Algorithmus mit quadratischer Laufzeit, der das bestmögliche (=günstigste) aller Alignments findet

Länge-basiertes Satzalignment: Gale & Church (1993)

- Korpus
  - 15 Wirtschaftsberichte der Union Bank of Switzerland (UBS-Korpus)
  - Originalsprache: Deutsch
  - Übersetzungen: Englisch und Französisch
  - Umfang (Englisch): 14680 Wörter, 725 Sätze, 188 Absätze
  - Gegeben: Alignment der Absätze
- Grundidee:
  - Hohe Korrelation zwischen der Anzahl der Zeichen eines Absatzes in den verschiedenen Sprachen (Korrelationskoeffizient Englisch-Deutsch: 0,991)

  → Kurze Sätze werden in kurze Sätze übersetzt, lange Sätze werden in lange Sätze übersetzt.
  - Also: Finde ein Alignment, so dass innerhalb jedes beads die Anzahl der Zeichen aus beiden Sprachen möglichst gleich ist.
  - Außerdem werden 1:1 Korrespondenzen bevorzugt.
    (Beschränkung der möglichen Korrespondenzen auf: 1:1, 1:0, 0:1, 2:1, 1:2, 2:2)

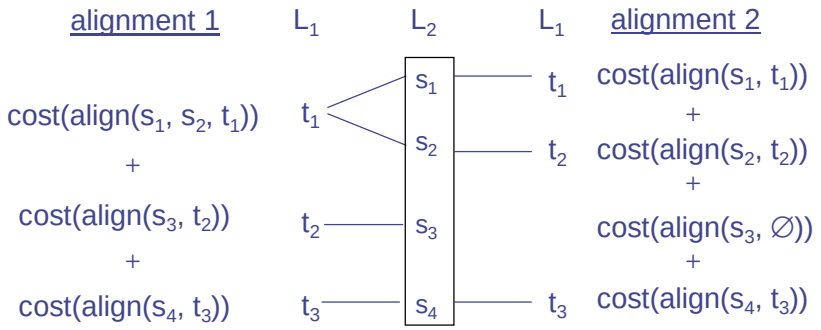
- Gegebenes Absatz-Alignment:
  - halbautomatisches Alignment der Absätze: Korrektur von Hand (nur möglich, da das Korpus relativ klein ist)
  - Ohne vorheriges Absatz-Alignment steigt die Fehlerrate um das dreifache (siehe Folie 9).
  - Aber man kann den Algorithmus prinzipiell auch zweimal laufen lassen: einmal zum Absatz-Alignment, dann zum Satz-Alignment. Dadurch sind auch größere Korpora (wie z.B. die kanadischen Parlamentsberichte) verarbeitbar.
- Korrelationskoeffizient: 1 = perfekt korreliert, 0 = keinerlei Korrelation

Gale & Church: Algorithmus

- Ziel: Finde für die parallelen Absätze S und T das Alignment mit den geringsten "Kosten" (und damit der höchsten Wahrscheinlichkeit):

  arg min cost(align(s_1,…,s_i,t_1,…,t_j))

- Dazu werden die parallelen Absätze in alle möglichen Sequenzen von beads eingeteilt, und die jeweiligen Kosten addiert:



- alignment 1 und alignment 2 sind nur zwei von vielen Möglichkeiten, Absatz S (bestehend aus den Sätzen s_1 bis s_4) und Absatz T (bestehend aus den Sätzen t_1 bis t_3) in beads einzuteilen.
- alignment 1 gruppiert s_1 und s_2 mit t_1 (2:1), s_3 mit t_2 (1:1), und s_4 mit t_3 (1:1).
- alignment 2 gruppiert s_1 mit t_1 (1:1), s_2 mit t_2 (1:1), s_3 bildet allein einen bead (1:0), und s_4 wird mit t_3 gruppiert (1:1).

- Die Gesamtkosten eines alignments ergeben sich durch die Addition der Kosten der einzelnen beads (wie die Kosten eines einzelnen beads berechnet werden, steht auf Folie 8)

Kostenberechnung der einzelnen beads: abhängig von

- l_1 und l_2 (= Anzahl der Zeichen von S und T in dem jeweiligen bead)
- der Korrespondenzkategorie (match)

$$\begin{aligned}\text{cost}(l_1,l_2,\text{match}) &= -\log P(\text{match} \mid \text{diff}(l_1,l_2)) \\ &\approx -\log (P(\text{match}) \cdot P(\text{diff}(l_1,l_2) \mid \text{match})) \\ &\approx -\log (P(\text{match}) \cdot (1 - P(\text{diff}(l_1,l_2))))\end{aligned}$$

$$\text{diff}(l_1,l_2) = \frac{|(l_2 - l_1)|}{\sqrt{l_1 \cdot 6{,}8}}$$

- Definition von $\text{diff}(l_1,l_2) \rightarrow$ je ähnlicher l_1 und l_2, desto kleiner $\text{diff}(l_1,l_2)$
- Definition von $P(\text{diff}(l_1,l_2)) \rightarrow$ je kleiner $\text{diff}(l_1,l_2)$, desto kleiner $P(\text{diff}(l_1,l_2))$
- je kleiner $P(\text{diff}(l_1,l_2))$, desto größer $(1 - P(\text{diff}(l_1,l_2)))$

$\Rightarrow$ ***je ähnlicher l_1 und l_2, desto größer $(1 - P(\text{diff}(l_1,l_2)))$***

Korrespondenz-Kategorie	Häufigkeit	**P(match)**
1:1	1167	0,89
1:0 oder 0:1	13	0,0099
2:1 oder 1:2	117	0,089
2:2	15	0,011
	1312	1,00

$\Rightarrow$ ***je häufiger eine Korrespondenz-Kategorie im UBS-Korpus, desto größer P(match)***

- cost(l_1,l_2,match)
  - - log macht aus der Wahrscheinlichkeit (zwischen 0 und 1) Kosten (zwischen unendlich und 0)
  - erste Umwandlung: nach dem Satz von Bayes (der konstante Faktor wird weggelassen)
  - zweite Umwandlung: vereinfachte Variante der Formel bei Gale & Church (1993), S. 83
- diff (l_1,l_2)
  - Je größer die Anzahl der Zeichen, desto größer auch die durchschnittliche Varianz der absoluten Differenz $|l_1 - l_2|$.
  - Daher wird die absolute Differenz noch mit einem Wert normalisiert, der von der Satzlänge abhängt.
- Definition von P(diff (l_1,l_2))
  - Tatsächliche Definition von P(diff (l_1,l_2)): s. S. 83/84 bei Gale & Church (dortige Bezeichnung: Prob(|δ|))
  - Aus dieser Definition lässt sich folgendes ableiten:
    - je kleiner diff (l_1,l_2), desto kleiner P(diff (l_1,l_2))
    - $0 \leq P(\text{diff}(l_1,l_2)) \leq 1$
- Um P(match) der jeweiligen Korrespondenz-Kategorie zu ermitteln,
  - wurden die Sätze des Korpus (= 15 Berichte der UBS) von 3 „menschlichen“ Alignern einander zugeordnet,
  - und dann gezählt, wie häufig jede Korrespondenzkategorie tatsächlich aufgetreten ist (dabei wurden die drei Fälle von „komplizierteren“ Korrespondenzen (3:1 und 3:2) nicht berücksichtigt).
  - P(match) = Häufigkeit der jew. Kategorie / Häufigkeit aller berücksichtigten Kategorien (z.B. 1167/1312 = 0,89 für die Kategorie 1:1)

Gale & Church: Evaluation

Fehlerraten

- insgesamt: 4,2%
- Sprachpaare:

Englisch-Französisch	Englisch-Deutsch
5,8%	2,7%

- Korrespondenzkategorien:

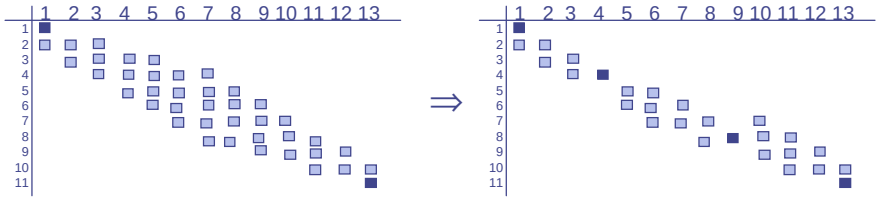
1:0	1:1	2:1	2:2	3:1	3:2
100%	2%	9%	33%	100%	100%

- Auswahl der Alignments (die besten 80%): 0,7%
- Kanadische Parlamentsberichte: 2%
- Satzlänge in Anzahl von Wörtern: 6,5%
- ohne vorheriges Absatz-Alignment: 12,9%

- Achtung: Gale & Church geben nur die Fehlerraten für die Sprachpaare Engl.-Franz. und Engl.-Dt. an.
- Zur Evaluation wurde das Alignment des Algorithmus mit dem „menschlichen“ Alignment verglichen:
  - 3 menschliche Aligner: Englisch-, Französisch-, bzw. Deutsch-Muttersprachler
  - Überlappungskriterium: mindestens ein Nebensatz gleich
- Sprachpaare: Aus der Angabe, dass die Fehlerrate beim Sprachpaar Engl.-Franz. höher ist als beim Sprachpaar Engl.-Dt., lässt sich nicht ableiten, dass das Alignment von Englisch und Französisch generell schwieriger ist. Die höhere Fehlerrate liegt in diesem Fall viel eher daran, dass beim UBS-Korpus sowohl die französischen als auch die englischen Texte Übersetzungen der deutschen Originaltexte sind.
- Korrespondenzkategorien:
  - Die 1:0 Korrespondenzen (13 Fälle von 1315) werden nie richtig erkannt.
  - Die Kategorien 3:1 und 3:2 (3 Fälle von 1315) werden deshalb nie erkannt, weil der Algorithmus sie von vornherein ausgeschlossen hat (s. Folie 8)
- Wenn man die Alignments nach ihren Kosten bewertet, und nur die günstigsten (=besten) 80% evaluiert, fällt die Fehlerrate auf 0,7%.
- Das Alignment der kanadischen Parlamentsberichte (Canadian Hansards) scheint einfacher zu sein als das Alignment des UBS-Korpus. Das mag daran liegen, dass die Parlamentsberichte wortgetreuer übersetzt wurden.
- Wenn man die Satzlänge in Anzahl von Wörtern (statt in Anzahl der Zeichen) misst, erhöht sich die Fehlerrate auf 6,5%. Das mag u.a. daran liegen, dass deutsche Komposita im Englischen häufig durch mehrere Wörter übersetzt werden (z.B. Fluglotse = air traffic controller), und die Anzahl von Wörtern daher kein so gutes Maß ist.
- Ohne vorheriges Absatz-Alignment (entweder halbautomatisch mit menschlicher Korrektur, oder durch zweimalige Anwendung des Algorithmus, s. Folie 6), erhöht sich die Fehler um das Dreifache!

Satzalignment mit Lexikalischen Methoden (Kay & Röscheisen 1993)

- Robustes Satzalignment durch Verwendung lexikalischer Information
- Algorithmus:
  - Annahme: Alignment der jeweiligen ersten und letzten Sätze der beiden Texte → Anfangs- und Endanker
  - Dann (wiederholen bis alle Sätze zugeordnet sind):
    1) Alle möglichen Alignments festlegen (Hüllkurve in Kissenform).
    2) Wortpaare auswählen, die ähnliche Verteilungen aufweisen und häufig genug vorkommen.
    3) Satzpaare finden, die viele der ausgewählten Wortpaare enthalten.
    4) Die besten Satzpaare als zusätzliche Anker auswählen und Schritte 1-4 wiederholen.



- Fehlerraten: 4% (Scientific American), 0,7% (Kanadische Parlamentsberichte)

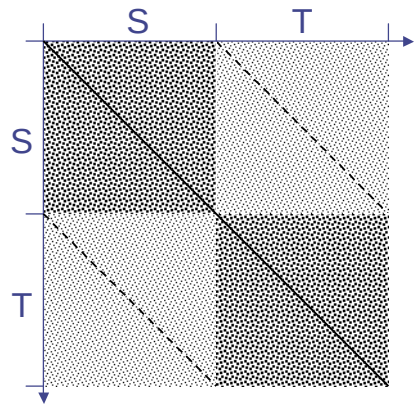
- Grundideen:
  - Aus dem partiellen Wortalignment (nur einige ausgewählte Wörter werden einander zugeordnet) wird das Satzalignment abgeleitet.
  - Das partielle Wortalignment basiert auf der Annahme, dass zwei Wörter einander zugeordnet werden können, wenn ihre statistische Verteilung sehr ähnlich ist.
- Nachteil: Das Satzalignment nach Kay & Röscheisen ist nicht sehr effizient/schnell.
- Außerdem sind Umstellungen (2:2) und Löschungen (1:0) sehr problematisch.
- Grafik: Die jeweiligen Endanker sind mit dunklen Kästchen dargestellt, die restlichen festgelegten Alignments mit hellen Kästchen.

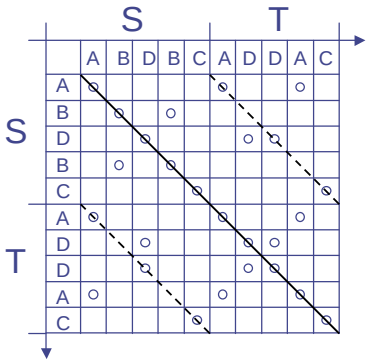
Offset Alignment durch Signalverarbeitungstechniken (Church 1993)

- Motivation: Satzlängen-basierte Methoden versagen bei “fehlerhaften” Texten (OCR, markup)
- Daher: Alignment von Offset-Positionen statt Satzalignment (Zuordnung der Sätze in beads)
- Ziel: identische Zeichenketten finden (verwandte Wörter, Eigennamen, Zahlen) → Offset-Positionen
- Methode:
  - Original und Übersetzung konkatenieren
  - dot plot des konkatenierten Textes erstellen (identische Zeichenketten werden durch einen Punkt markiert)
  - Heuristische Suche entlang der Diagonalen → Offset-Positionen finden

- OCR = optical character recognition (z.B. wenn man einen gedruckten Text einscannt)
- Offset-Positionen: beliebige Positionen im Text
- Länge der verglichenen Zeichenketten: 4 Zeichen
- Bei sehr unterschiedlichen Sprachen gibt es zwar keine verwandten Wörter, aber Eigennamen und Zahlen sind gleich (bei anderen Schriftsystem, wie z.B. Chinesisch, aber möglicherweise noch nicht einmal das)

Dot Plots





- Die jeweiligen Texte sind natürlich sich selbst ähnlicher als dem anderen Text. Daher gibt es in den Quadraten, in denen die Texte mit sich selbst verglichen werden, viel mehr Übereinstimmungs-punkte.
- Interessanter sind die dünnen Diagonalen in den helleren Feldern: Die Punkte entlang dieser Diagonalen sind die gesuchten Offset-Positionen.

Wortalignment

- Motivation: Erstellung probabilistischer bilingualer Lexika
- Methode:
  - Textalignment → Wortalignment
    - χ^2–Test
    - 1:1 Korrespondenz
  - Auswahlkriterium: Häufigkeit
- Ausblick: existierende bilinguale Lexika einbeziehen

- Wie wir bei der Methode von Gale & Church gesehen haben, lässt sich derselbe Algorithmus sowohl zum Absatz-Alignment als auch zum Satz-Alignment verwenden.
- Um probabilistische Lexika zu erstellen (s. Folie 2), wird das Alignment nun auf Wortebene durchgeführt.
- Mit Hilfe des χ^2–Tests (vgl. Manning & Schütze, Kap. 5.3.3) und der Einschränkung auf 1:1-Korrespondenzen gelangt man zu verlässlichen Ergebnissen.
- Er werden jedoch nur die Einträge in das Lexikon aufgenommen, die genügend häufig vorkommen.
- Beim Wortalignment bietet sich an, nicht nur statistische Methoden zu verwenden, sondern auch bestehende bilinguale Lexika als linguistische Wissensquelle mit einzubeziehen.

Literatur

- Manning, C. & Schütze, H. (1999): *Foundations of Statistical Natural Language Processing*. MIT Press, Kap. 13.1 und 13.2.
- Gale, W.A. & Church, K.W. (1993): A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19, S. 75-102.
- Kay, M. & Röscheisen, M. (1993): Text-translation alignment. *Computational Linguistics* 19, S. 121-142.
- Church, K.W. (1993): Char_align: A program for aligning parallel texts at the character level. In: *ACL 31*, S. 1-8.

- Gale & Church (1993) ist online verfügbar (im PDF-Format oder als ps-Datei): http://citeseer.nj.nec.com/gale93program.html
- Church (1993) findet man hier: http://citeseer.nj.nec.com/35831.html